Raymond Mejia

Neural Net Models to Characterize Integral Membrane Proteins - Preliminary Report

We describe a class of models that are designed to help identify extracellular, membrane and cytosolic protein segments, and the direction in which an integral membrane protein traverses (or simply penetrates) the cell membrane.

1. Introduction

Neural nets have been used for many learned tasks - for perceptrons Minsky and Papert [6], parallel processing Rumelhart, Hinton and McClelland [8], prediction of secondary structure of proteins Qian and Sejnowski [7], and prediction of membrane-spanning protein topology Knepper [2]. We describe a class of models to determine extracellular, cytosolic and membrane domains from the amino acid sequence of integral membrane proteins.

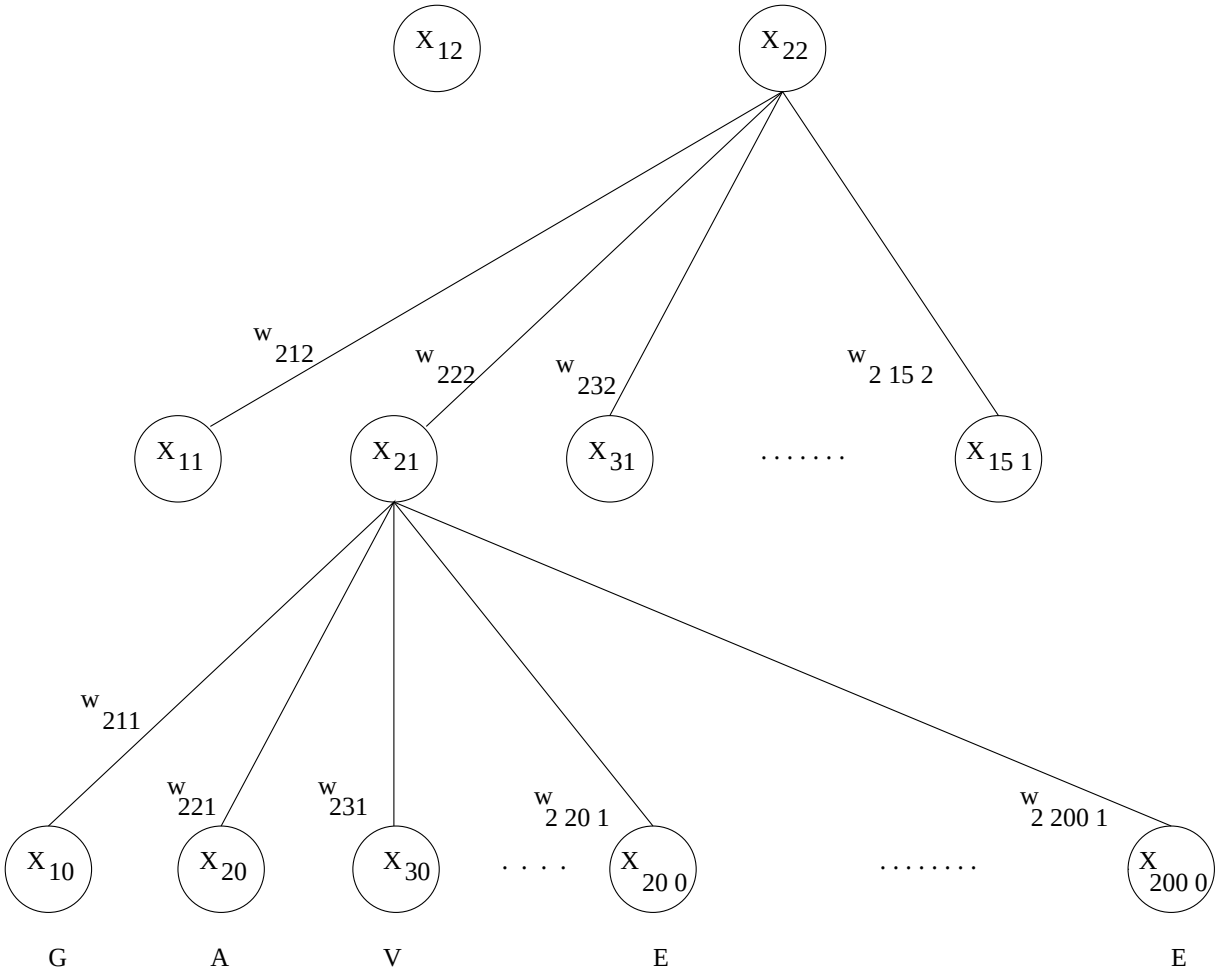

Figure 1: This diagram shows a model with an input window 10 amino acids wide, one hidden layer with 15 nodes, and two output units. A protein is read from the H- to C-terminus by shifting it through the window one position at a time. Each node at a given level is connected to all nodes at the prior, lower level. For clarity, connections that enter are shown for only one node per level.

2. Model

In order to identify the location of segments of a protein, we use a feed-forward network of non-linear units with one hidden layer and adjustable weights as follows:

$$\begin{aligned} x_{ik} &= \textstyle\sum_j w_{ijk} X_{jk-1}, \\ X_{ik} &\equiv f(x_{ik}) = 1/[1+\exp^{-x_{ik}}], \end{aligned} \tag{1}$$

where X_{ik} is the output of node i at level k; $k = 1, 2$; w_{ijk} is the weight connecting node i at level k to node j at level $k-1$; x_{ik} is the input to node i at level k; and f is the activation function, taken here to be a logistic.

As the network is trained, the weights are adjusted, and we minimize the error E, $E = \sum_p E^p$, where

$$E^p = \frac{1}{2}\sum_i [X^p_{i2} - T^p_i]^2, \tag{2}$$

and $\mathbf{T}^p$ is the target vector for a given input pattern. Henceforth, we will omit the superscript p for brevity.

We use steepest descent to maximize the descent of E along the gradient, and use the chain rule to write

$$\frac{dE}{dw_{ij2}} = \frac{dE}{dX_{i2}}\frac{dX_{i2}}{dx_{i2}}\frac{dx_{i2}}{dw_{ij2}}. \tag{3}$$

The last term in (3) depends on $\mathbf{X_1}$, the output of the hidden layer. With $f \in C^1$, use of (3) permits back-propagation of the error and correction of $w_{ij2}, \forall i, j$ as follows:

$$\Delta w_{ij2} = \eta(T_i - X_{i2})X_{i2}(1 - X_{i2})X_{j1}, \tag{4}$$

where η, $0 < \eta \leq 1$, is the learning rate.

The weights in the hidden layer are updated similarly, with a correction Δw_{ij1} written as:

$$\Delta w_{ij1} = (1 - X_{i1})X_{j0}\sum_l \Delta w_{li2} w_{li2}. \tag{5}$$

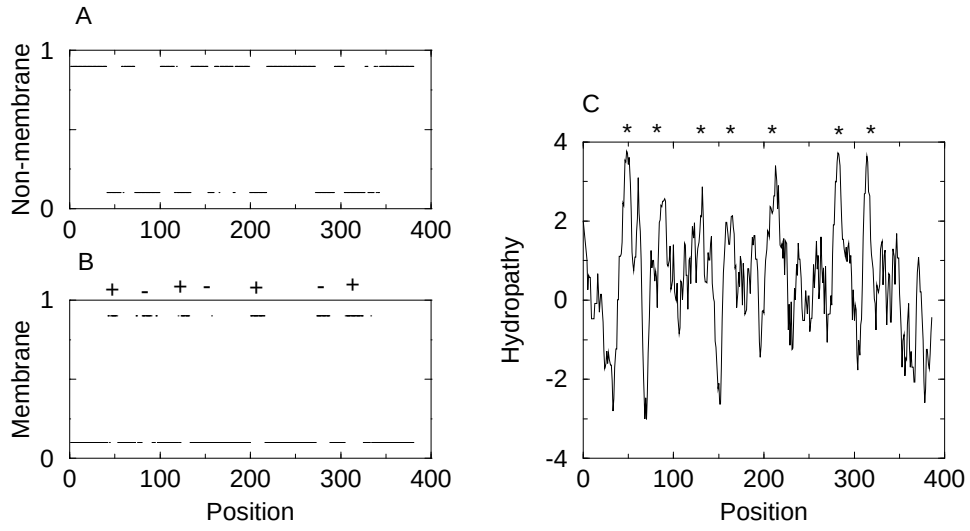

Figure 2: Output for the human oxytocin receptor. Panel A shows model output for the non-membrane node. A value of 0.9 identifies a non-membrane position; a value of 0.1 indicates the position in the protein is not identified as non-membrane (which does not necessarily mean that it will be identified as in the membrane). Panel B shows output for the membrane node. 0.9 = membrane; 0.1 = not in the membrane; + identifies a protein segment in the membrane that is oriented toward the cytosole when read from the N- toward the C-terminus; − indicates a membrane segment is exiting the cell. Panel C shows the hydropathy index [4]. Positive values indicate hydrophobicity and negative values hydrophilicity. Membrane segments are marked by an $*$.

Thus, to train a net we solve equations (1), and if the error is not less than a prescribed tolerance, we use equations (4) and (5) to obtain new weights, $\mathbf{w}_k^{new} = \mathbf{w}_k^{old} + \Delta\mathbf{w}_k$, $k = 1, 2$.

Models are designed to identify the location of protein segments with respect to cell structure (extracellular, in the cell membrane, intracellular). These include models to discern: membrane or non-membrane; pericellular or cytosole; pericellular, membrane or cytosole; pericellular, membrane inward, cellular or membrane outward domains.

Analysis of a protein with a particular model simply requires solution of equations (1) with a given set of weights.

3. Results

Fig. 1 illustrates such a model with two output units that identify membrane and non-membrane segments, respectively. The protein is read from N-terminus to C-terminus through a window of ten amino acid positions, and shifted one position until the entire polypeptide is read. For any window, the input vector $\mathbf{X}_0^p$ consists of 200 nodes, one for each of the 20 amino acids that may be present at each position on the protein; $\mathbf{X}_0^p = \{X_{j0}^p : X_{j0}^p = 0, 1; \quad j = 1, ..., 200\}$. Presence of an amino acid is signified by a 1 in the appropriate position, absence by a 0, so that an input window has 20 ones and 180 zeros. In addition, the model has a hidden layer with fifteen nodes and the two output nodes.

The neural network is trained with a set of proteins whose structure has been identified. For this study, we have trained a network with sixteen G-protein coupled receptors that have been characterized using crystallography and hydrophobicity to identify extracellular, membrane and intracellular domains.

Fig. 2 shows regions outside the membrane (Panel A), in the membrane (Panel B), and the hydropathy index [4] (Panel C) for the human oxytocin receptor [3]. Amino acid positions are numbered from the H-terminus to the C-terminus of the protein. The output has been filtered through a window that is five positions wide and with a threshold of 0.6, so that a mean value ≥ 0.6 yields 0.9 and a value < 0.6 yields 0.1. Seven membrane spanning regions are identified and labeled + when entering and − when exiting the cell. (Direction has been determined with a three node model that distinguishes extracelullar, membrane and cytosolic domains.) These results are consistent with the motif for a G-protein coupled receptor [9]. Membrane segments identified by the model are also labeled in the hydropathy graph, where large positive values indicate hydrophobicity and large negative values indicate hydrophilicity.

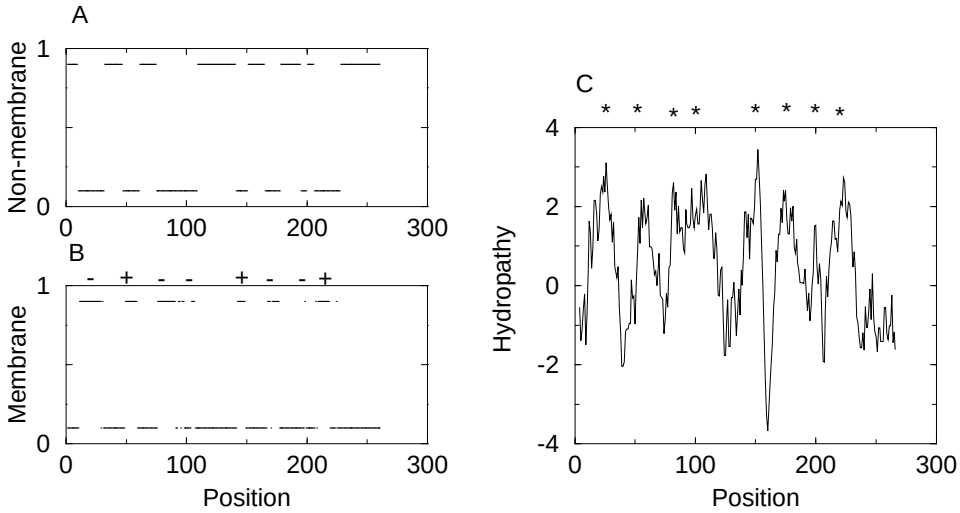

Figure 3: Output for the rat chip28 water channel [5]. See Fig. 2 for a description of each panel.

Fig. 3 shows the membrane and non-membrane positions for the rat chip28 water channel [5], as well as a hydropathy map. This shows that chip28 conforms to the aquaporin motif [1]. Note the two segments that loop into but do not span the membrane (Panel B).

The vasopressin-regulated urea transporter of rabbit [10] is characterized as shown in Fig. 4. Ten membrane domains are identified and oriented in Panel B and marked for comparison in Panel C.

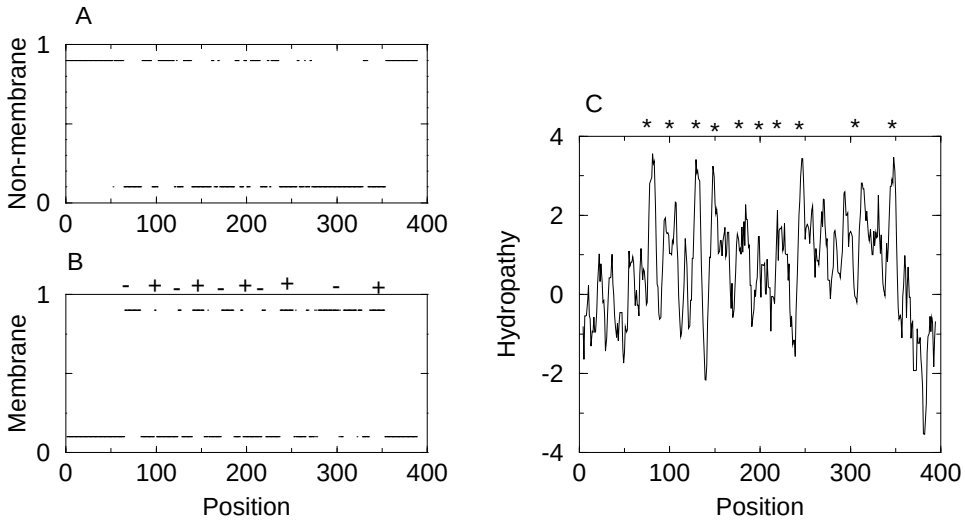

Figure 4: Output for the vasopressin-regulated urea transporter [10]. See Fig. 2 for a description of each panel.

4. Conclusion

We have described neural net models that aid in the characterization of integral membrane proteins. The time required to train a network varied from minutes to several hours using an IBM RS/6000 Model 370 workstation, depending on the number of output nodes (2 - 4) used. The time to analyze each protein was less than a minute.

Acknowledgements

The results presented in this paper are based on joint work with Dr. Mark A. Knepper, at the National Institutes of Health, Bethesda, MD that was motivated by work conducted by Ross A. Knepper [2].

5. References

1 Jung, J.S., Bhat, R.V., Preston ,G.M., Guggino, W.B., Baraban, J.M., Agre, P.: Molecular characterization of an aquaporin cDNA from brain: candidate osmoreceptor and regulator of water balance. Proc. Natl. Acad. Sci. U.S.A. **91** (1994), 13052-13056.

2 Knepper, R.A.: Identification of membrane-spanning domains of integral membrane proteins using a neural network. Washington STS Competition. (1994).

3 Kimura, T., Tanizawa, O., Mori, K.: Structure and expression of a human oxytocin receptor. Nature **356** (1992), 526-529.

4 Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157** (1982), 205–132.

5 Li, J., Nielsen, S., Dai, Y., Lazowski, K.W., Christensen, E.I., Tabak, L.A., Baum, B.J.: Examination of rat salivary glands for the presence of the aquaporin CHIP. Pflugers Arch **428** (1994), 455–60.

6 Minsky, M., Papert, R.: Perceptrons. MIT Press, Cambridge, MA (1969).

7 Qian, N., Sejnowski, T.J.: Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. **202** (1988), 865–884.

8 Rumelhart, D.E., Hinton, G.E., McClelland, J.L.: A general framework for parallel distributed processing. in Parallel Distributed Processing. MIT Press, Cambridge, MA **1** (1986), 45–76.

9 Voet, D., Voet J.G.: Biochemistry. John Wiley & Sons, New York, NY (1995), 1268.

10 You, G., Smith, C., Kanai, Y., Lee, W., Stelzner, M., Hediger, M.A.: Cloning and characterization of the vasopressin-regulated urea transporter. Nature **365** (1993), 844–847.

Addresses: Raymond Mejia, Laboratory of Kidney and Electrolyte Metabolism, NHLBI and Mathematical Research Branch, NIDDK, National Institutes of Health, Bethesda, MD 20892–2690, USA.
email: ray@helix.nih.gov